PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2003-316589

(43)Date of publication of application : 07.11.2003

(51)Int.Cl. G06F 9/46
G06F 3/06

(21)Application number : 2002-120216

(22)Date of filing : 23.04.2002

(71)Applicant : HITACHI LTD

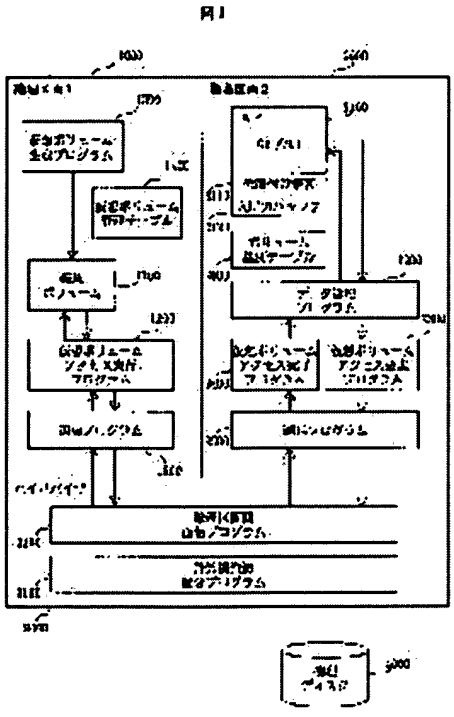
(72)Inventor : NAGASUGA HIROFUMI
KIYOI MASAHIRO
OTSUJI AKIRA
IKEGAYA NAOKO
HIRAIWA YURI

(54) REAL MEMORY USING METHOD

(57)Abstract:

PROBLEM TO BE SOLVED: To solve the problem that delay due to the concentration of access is generated, or that the change of a user program is necessary for adopting a method for preventing the generation of the delay in a system where a shared data is disposed in an external storage device.

SOLUTION: A selected arbitrary computer in a virtual computer system is provided with a step for converting an address designated in an input/output request command transmitted from another computer to an external storage device by a communicating means into an address in a corresponding data region in a real storage device held in the selected logical block, a step for, when the command is an input request, returning the data stored in the data region of the converted address to a computer being the origin of transmission of the command, and a step for, when the command is an output request, storing the data transmitted with the command in the data region of the converted address.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開2003－316589
（P2003－316589A）
(43)公開日　平成15年11月7日(2003.11.7)

(51)Int.Cl.[7]		識別記号	FI		テーマコード(参考)
G06F	9/46	350	G06F	9/46	350　5B065
	3/06	301		3/06	301J　5B098
					301K

審査請求　未請求　請求項の数８　ＯＬ　（全 14 頁）

(21)出願番号　特願2002－120216(P2002－120216)

(22)出願日　平成14年４月23日(2002.4.23)

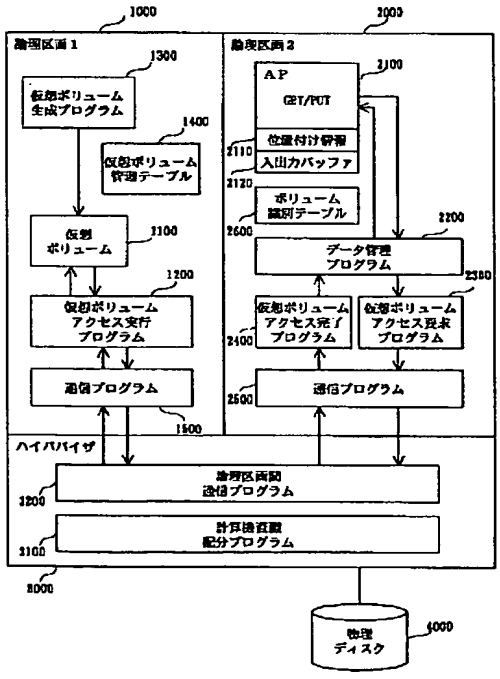
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　長須賀　弘文
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(72)発明者　清井　雅広
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(74)代理人　100075096
弁理士　作田　康夫

最終頁に続く

(54)【発明の名称】　実記憶利用方法

(57)【要約】　（修正有）
【課題】外部記憶装置に共用データを配置しているシステムでは、アクセスの集中による遅延の発生や、この遅延発生を回避する手法を採用するために発生するユーザプログラムの変更の必要性といった課題があった。
【解決手段】　仮想計算機システムにおいて、選択した任意の計算機には、他の計算機から、通信手段によって送られた外部記憶装置への入出力要求コマンドの中で指定されたアドレスを、当該選択された論理区画が保持する実記憶装置内の対応するデータ域のアドレスに変換するステップを設ける。そして、そのコマンドが入力要求の時は、変換したアドレスのデータ域に格納されているデータをコマンドの送信元である計算機へ返送するステップを設ける。さらに、コマンドが出力要求の時は、コマンドと一緒に送られてきたデータを、変換したアドレスのデータ域に格納するステップを設ける。

図１



【特許請求の範囲】

【請求項１】物理的記憶装置を論理的に分割した論理区画ごとに記憶されたプログラムを実行させることにより、複数の計算機を仮想的に実行することができる仮想計算機システムにおいて、

前記論理区画のうち少なくとも１つの第１の論理区画に記憶されたプログラムの実行を選択する選択手段とを有し、

少なくとも前記第１の論理区画には、

第２の論理区画との間で通信可能な通信手段と、

前記物理的記憶装置に対するデータの入力要求コマンドを受信する受信手段と、

前記入力要求コマンドにおいて指定されるデータアドレスを、前記第１の論理区画に割当てられた論理的データアドレスに変換する変換手段と、

前記論理的データアドレスに基づいて、前記物理的記憶装置に対して前記データを格納する格納手段とを有することを特徴とする仮想計算機システム。

【請求項２】請求項１に記載の仮想計算機システムにおいて、

少なくとも前記第１の論理区画には、

さらに、前記選択手段により選択された第１の論理区画に記憶されたプログラムを実行することにより発行された、前記第２の論理区画に対するデータの入力要求に基づいて、前記物理的記憶装置に対する入力要求コマンドを生成する生成手段とを有し、

前記通信手段は、

前記入力要求コマンドと前記入力要求コマンドを用いて第２の論理区画に対して格納するデータとを、前記第２の論理区画に対して送信することを特徴とする仮想計算機システム。

【請求項３】物理的記憶装置を論理的に分割した論理区画ごとに記憶されたプログラムを実行させることにより、複数の計算機を仮想的に実行することができる仮想計算機システムにおいて、

前記論理区画のうち少なくとも１つの第１の論理区画に記憶されたプログラムの実行を選択する選択手段とを有し、

少なくとも前記第１の論理区画には、

前記第２の論理区画との間で通信可能な通信手段と、

前記物理的記憶装置に対するデータの出力要求コマンドを受信する受信手段と、

前記出力要求コマンドにおいて指定されるデータアドレスを、前記第１の論理区画に割当てられた論理的データアドレスに変換する変換手段とを有し、

前記通信手段は、

前記論理的データアドレスに基づいて、前記物理的記憶装置に格納されたデータを、前記入力要求コマンドを発行した第２の論理区画に対して送信することを特徴とする仮想計算機システム。

【請求項４】請求項３に記載の仮想計算機システムにおいて、

少なくとも前記第１の論理区画には、

さらに、前記選択手段により選択された第１の論理区画に記憶されたプログラムを実行することにより発行された、第２の論理区画に対するデータの出力要求に基づいて、前記物理的記憶装置に対する出力要求コマンドを生成する生成手段とを有し、

前記通信手段は、

前記出力要求コマンドを前記第２の論理区画の通信手段に対して送信することを特徴とする仮想計算機システム。

【請求項５】物理的記憶装置を論理的に分割した論理区画ごとに記憶されたプログラムを実行させることにより、複数の計算機を仮想的に実行することができる仮想計算機システムによる実記憶利用方法において、

前記論理区画のうち少なくとも１つの第１の論理区画に記憶されたプログラムの実行を選択する選択ステップと、

少なくとも前記第１の論理区画において、

第２の論理区画との間で通信可能な通信ステップと、

前記物理的記憶装置に対するデータの入力要求コマンドを受信する受信ステップと、

前記入力要求コマンドにおいて指定されるデータアドレスを、前記第１の論理区画に割当てられた論理的データアドレスに変換する変換ステップと、

前記論理的データアドレスに基づいて、前記物理的記憶装置に対して前記データを格納する格納ステップとを有することを特徴とする実記憶利用方法。

【請求項６】物理的記憶装置を論理的に分割した論理区画ごとに記憶されたプログラムを実行させることにより、複数の計算機を仮想的に実行することができる仮想計算機システムによる実記憶利用方法おいて、

前記論理区画のうち少なくとも１つの第１の論理区画に記憶されたプログラムの実行を選択する選択ステップとを有し、

少なくとも前記第１の論理区画において、

前記第２の論理区画との間で通信可能な通信ステップと、

前記物理的記憶装置に対するデータの出力要求コマンドを受信する受信ステップと、

前記出力要求コマンドにおいて指定されるデータアドレスを、前記第１の論理区画に割当てられた論理的データアドレスに変換する変換ステップとを有し、

前記通信ステップは、

前記論理的データアドレスに基づいて、前記物理的記憶装置に格納されたデータを、前記入力要求コマンドを発行した第２の論理区画に対して送信することを特徴とする実記憶利用方法。

【請求項７】複数の計算機ノードと、該複数の計算機ノ

ードが共有する物理記憶装置とを有する計算機システムにおいて、
前記のうち少なくとも１つの第１の計算機ノードを選択する選択手段とを有し、
少なくとも前記第１の計算機ノードには、
第２の計算機ノードとの間で通信可能な通信手段と、
前記物理的記憶装置に対するデータの入力要求コマンドを受信する受信手段と、
前記入力要求コマンドにおいて指定されるデータアドレスを、前記第１の論理区画に割当てられた論理的データアドレスに変換する変換手段と、
前記論理的データアドレスに基づいて、前記物理的記憶装置に対して前記データを格納する格納手段とを有することを特徴とする計算機システム。
【請求項８】複数の計算機ノードと、該複数の計算機ノードが共有する物理記憶装置とを有する計算機システムにおいて、
前記のうち少なくとも１つの第１の計算機ノードを選択する選択手段とを有し、
少なくとも前記第１の計算機ノードは、
前記第２の計算機ノードとの間で通信可能な通信手段と、
前記物理的記憶装置に対するデータの出力要求コマンドを受信する受信手段と、
前記出力要求コマンドにおいて指定されるデータアドレスを、前記第１の計算機ノードに割当てられた論理的データアドレスに変換する変換手段とを有し、
前記通信手段は、
前記論理的データアドレスに基づいて、前記物理的記憶装置に格納されたデータを、前記入力要求コマンドを発行した第２の計算機ノードに対して送信することを特徴とする計算機システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】大容量の実記憶装置の利用方法に関し、特に複数の計算機を仮想的に実行することができる仮想計算機システムにおける実記憶装置の利用方法に関する。
【０００２】
【従来の技術】近年の基幹系のサーバシステムでは、膨大するデータ量や処理量に柔軟に対応することを可能にするため、複数の計算機イメージから構成されるクラスタ型のシステム構成を採用するのが一般的である。さらに、多くの計算機システムでは、各計算機イメージ間で共用が必要なデータは、各計算機イメージからアクセスできる共用の外部記憶装置に配置している。
【０００３】しかし、共用データに対しては、複数の計算機システムからのアクセス要求が発生するため、データの整合性を保証するため、排他処理を実施ながらデータアクセスを制御する必要がある。一般的には、こうした排他処理用に利用する制御情報も、共用の外部記憶装置に配置する。そのため、システムの応答性能やスループットの向上を図るためには、外部記憶装置に配置した排他処理用のデータに高速にアクセスできる環境が必要である。
【０００４】しかし、外部記憶装置内では、ディスクを回転させながら、対象となる領域へアクセスするため、共用の外部記憶装置に対してアクセスが集中した場合は、ディスクの回転待ちが、アクセスの遅延を発生させる要因となる。
【０００５】それを解決するための手段として特開平９－２９３０５５がある。特開平９－２９３０５５では、共用の外部記憶装置の代わりに拡張メモリを配置し、このメモリ域に共用ファイルを配置すると共に、排他処理で利用するロック領域を設けている。こうすることで、共用するデータへのアクセス時の回転待ち等によるアクセス遅延を回避すると共に、制御情報へのアクセスの高速化により、排他処理の効率化を図っている。
【０００６】
【発明が解決しようとする課題】しかし、上記従来技術を用いるためには、新たにアプリケーションプログラムを用意する必要性が生ずる。共用するデータを共用外部記憶装置に配置していたシステムでは、外部記憶装置へアクセスするようプログラムを作成していた。しかし、上記従来技術を利用するためには、それを、共用メモリへアクセスするように変更する必要がある。
【０００７】また、上述した課題を解決するための手法として、オペレーティングシステムが用意している外部記憶装置への入出力用インタフェースを用いて、上述した共用メモリにアクセスできる機能をオペレーティングシステムがサポートする手法を想定することができる。この手法では、外部記憶装置へアクセスするためにアプリケーションプログラムが発行した入出力要求を、オペレーティングシステムがトラップし、上記の共用メモリへのアクセスに変更する。そのため、各計算機ノードで動作しているオペレーティングシステムは、上記共用メモリに、外部記憶装置への入出力インタフェースを用いて、アクセスする機能をサポートする必要がある。オペレーティングシステムが、こうした機能をサポートするためには、膨大な開発量を要する。
【０００８】本発明の第１の目的は、アプリケーションプログラムに対しては、従来通りの外部記憶装置へアクセスするインタフェースでの動作を保証した上で、比較的少量のオペレーティングシステムの開発量で、共用データへのアクセスの高速化によるシステムの応答性能やスループットの向上を図る方法を提供することである。
【０００９】本発明の第２の目的は、こうしたデータへのアクセスに関して、別の論理区画域に配置させることで、データの機密性を向上させることである。
【００１０】

【課題を解決するための手段】上記課題を解決するために、論理分割により、ひとつの計算機システム上に複数の計算機イメージを構築することが可能なシステムにおいて、選択した任意の計算機イメージには、他の計算機イメージから、通信手段によって送られてきた外部記憶装置への入出力要求コマンド内で指定されたアドレスを、当該選択された論理区画が保持する実記憶装置内の対応するデータ域のアドレスに変換するステップを設ける。さらに、入出力要求コマンドが入力要求の時は、変換するステップにより求められたアドレスのデータ域に格納されているデータを入出力要求コマンドの送信元である計算機イメージへ返送し、入出力要求コマンドが出力要求の時は、入出力要求コマンドと一緒に送られてきたデータを、変換するステップにより求められたアドレスのデータ域に格納するステップを設ける。

【００１１】また、選択された計算機イメージ以外には、動作中のプログラムが発行したデータの入出力要求から、外部記憶装置への入出力要求コマンドを生成するステップを設ける。さらに、この入出力要求コマンドが、上記選択された計算機イメージのいずれかに格納されたデータの入力要求である場合には、当該入出力要求コマンドを、アクセス対象のデータが格納されている計算機イメージへ通信によって送り、入出力要求コマンドが選択された計算機イメージのいずれかへの出力要求である場合には、格納先の計算機イメージに、この入出力要求コマンドとデータを通信によって送るステップを設ける。

【００１２】なお、この解決手段は、互いに通信することが可能な複数の計算機ノードから構成されるクラスタ型計算機システムにおいても、適用することができる。

【００１３】

【発明の実施の形態】図１は、本発明の実施の形態を示す計算機システムの構成を示す。図１において、１０００、２０００はそれぞれ論理区画１、論理区画２である。また、３０００は、計算機システム内の資源を各論理区画に割当て、また、論理区画間で必要な通信を行うハイパバイザである。ハイパバイザ（３０００）によって、計算機システム内の資源を割当てられた論理区画１（１０００）及び論理区画２（２０００）に記憶されたプログラムを制御部（ＣＰＵ）により実行させることにより、仮想的に独立した複数の計算機として実行することができる。

【００１４】４０００は、計算機システムに接続された物理ディスク装置であり、本実施の形態では、ハイパバイザ（３０００）により、論理区画２（２０００）に割当てられている。

【００１５】１１００は、論理区画１（１０００）に割当てられた実記憶装置内に生成された仮想ボリュームである。この仮想ボリューム（１１００）を生成するのが、仮想ボリューム生成プログラム（１２００）である。

【００１６】１３００は、仮想ボリュームアクセス実行プログラムであり、論理区画２（２０００）から仮想ボリューム（１１００）に対するアクセス要求を受け取ったとき、その要求内容に応じて、仮想ボリューム（１１００）領域へのデータ書き込み、データ読み込み、または処理対象データ位置の計算を行う。

【００１７】１４００は、仮想ボリューム管理テーブルであり、仮想ボリューム（１１００）を管理するために必要な各種情報を格納する。

【００１８】１５００は、他の論理区画あるいはネットワークを介した外部の計算機との通信処理を実行する通信プログラムである。通信プログラム（１５００）は、論理区画１（１０００）から他の計算機に向けて通信を発信する処理と、他の計算機から論理区画１（１０００）への通信を受け付け、通信内容に応じて所定のプログラムを実行する処理を有する。通信内容が、仮想ボリューム（１１００）に対するアクセス要求である場合は、前述した仮想ボリュームアクセス実行プログラム（１２００）を起動する。また、仮想ボリュームアクセス処理が完了したあとに、その応答を要求元の論理区画１（１０００）へ送信するのも、通信プログラム（１５００）である。

【００１９】続いて、論理区画２の構成内容を説明する。２１００は論理区画２（２０００）内で実行されるアプリケーションプログラム（以下ＡＰ）である。ＡＰ（２１００）は、物理ディスク（４０００）、あるいは仮想ボリューム（１１００）への入出力処理の実行要求を含む。ＡＰ（２１００）内の２１１０は、入出力処理を実行する際、物理ディスク（４０００）あるいは仮想ボリューム（１１００）内での処理対象場所を示す位置付け情報である。また、２１２０は、入出力処理を行うデータを格納する入出力バッファである。

【００２０】２２００は、ＡＰ（２１００）からのデータ入出力要求を受付け、実際の入出力処理を実行するデータ管理プログラムである。

【００２１】２３００は、仮想ボリュームアクセス要求プログラムである。仮想ボリュームアクセス要求プログラム（２３００）は、ＡＰ（２１００）から要求されたデータ入出力の処理対象が仮想ボリュームであったときに、データ管理プログラム（２２００）によって呼び出される。そして、仮想ボリュームの管理元である論理区画１（１０００）に対する通信の実行を、通信プログラム（２５００）に要求する。

【００２２】２４００は、仮想ボリュームアクセス完了プログラムである。仮想ボリューム（１１００）に対するアクセスが完了すると、仮想ボリューム（１１００）の管理元である論理区画１（１０００）から、処理完了を示す応答が送信される。論理区画２（２０００）の通信プログラム（２５００）は、このアクセス完了の通信

を受け取ると、仮想ボリュームアクセス完了プログラム（２４００）を実行する。ＡＰ（２１００）からの入出力要求がデータ読込み要求の場合、仮想ボリュームアクセス完了プログラム（２４００）は、仮想ボリューム（１１００）から読み込んだデータを入出力バッファ領域に複写する。

【００２３】通信プログラム（２５００）は、論理区画１（１０００）内の通信プログラム（１５００）と同様に、他の論理区画あるいはネットワークを介した外部の計算機との通信処理を実行する。

【００２４】続いて、ハイパバイザ（３０００）の構成内容を説明する。３１００は、計算機資源配分プログラムである。計算機資源配分プログラム（３１００）は、実記憶装置、プロセッサといった計算機が有する資源を、論理区画１（１０００）及び論理区画２（２０００）に静的または動的に割当てる。これにより、論理区画１（１０００）、論理区画２（２０００）が、それぞれ独立した計算機システムイメージとして稼動できる環境を提供する。

【００２５】３２００は、ハイパバイザ（３０００）内の論理区画間通信プログラムである。論理区画間通信プログラム（３２００）は、論理区画１（１０００）から論理区画２（２０００）へ、またはその逆方向の通信要求がなされたとき、要求先の論理区画に通信データを伝達する。

【００２６】図２は、ＡＰ（２１００）から仮想ボリューム（１１００）に対する入出力要求が発生したときの、制御の流れを示す。データ管理プログラム（２２００）は、仮想ボリュームアクセス要求プログラム（２３００）を起動し、その後、通信プログラム（２５００）によって、論理区画１（１０００）へアクセス要求が発信される。この通信は、ハイパバイザ（３０００）の論理区画間通信プログラム（３２００）を経由して、論理区画１（１０００）内の通信プログラム（１５００）へ渡される。

【００２７】要求を受け取った論理区画１（１０００）側では、仮想ボリュームアクセス実行プログラム（１２００）を起動し、要求内容に従い仮想ボリューム（１１００）へのデータ書込み、あるいはデータ読込みを行ったのち、要求元の論理区画２（２０００）へ応答を返す。

【００２８】応答情報を受け取った論理区画２（２０００）内の通信プログラム（２５００）は、仮想ボリュームアクセス完了プログラム（２４００）を起動し、以下、データ管理プログラム（２２００）を介して、ＡＰ（２１００）の処理を再開する。

【００２９】続いて、図３を用いて、仮想ボリューム管理テーブル（１４００）の構成を説明する。仮想ボリューム管理テーブル（１４００）は、仮想ボリューム（１１００）を、物理ディスク（４０００）と同じインタフェースで利用するための構成情報と、各時点における仮想ボリューム（１１００）内でのアクセス場所を管理するための情報とから構成される。ここで、物理ディスク（４０００）内のデータをアクセスする場合、物理ディスク（４０００）内におけるデータの配置場所は、シリンダ番号、ヘッダ番号、及び、シリンダ番号とヘッダ番号とにより決定されるトラック内レコード番号によって特定される。各ボリュームのシリンダ数、ヘッダ数、トラック長は、装置毎に物理的に決められる。

【００３０】このような物理ディスク（４０００）へのアクセスと共通のインタフェースに対応するため、仮想ボリューム管理テーブル（１４００）は、仮想シリンダ数（１４１０）、仮想ヘッダ数（１４２０）、仮想トラック長（１４３０）を格納するフィールドを有する。これらのフィールドに格納された値により、仮想ボリューム（１１００）を構成するデータ領域の大きさが決定される。

【００３１】また、仮想ボリューム管理テーブル（１４００）内のレコード長（１４４０）は、ＡＰ（２１００）からアクセスするデータの最小単位を表すレコード長を格納する領域である。更に、データ先頭ポインタ（１４５０）は、仮想ボリューム（１１００）の領域を確保した実記憶装置内での先頭アドレスを格納するフィールドである。アクセスポインタ（１４６０）は、各時点においてアクセスすべき仮想ボリューム（１１００）内のデータ位置を示すアドレスを格納するフィールドである。

【００３２】仮想ボリューム管理テーブル（１４００）を構成する各フィールドの初期設定と、仮想ボリューム（１１００）データ領域の確保を行うのが、仮想ボリューム生成プログラム（１３００）である。図４は、仮想ボリューム生成部（１３００）が、仮想ボリューム管理テーブル（１４００）を初期設定した状態を表した図である。ここで、仮想シリンダ数（１４１０）が１０２４個、仮想ヘッダ数（１４２０）が１６個、仮想トラック長（１４３０）が３２７６８バイトである仮想ボリューム（１１００）が生成されている。また、各トラックに格納するデータ入出力の単位であるレコード長（１４４０）の初期値として、４０９６バイトを設定している。

【００３３】本実施の形態では、仮想ボリューム（１１００）を構成するデータ領域は、論理区画１（１０００）の実記憶装置内に、連続して確保される。データ配置順序は、シリンダ番号の昇順、及び各シリンダ内のトラック番号の昇順である。ただし、シリンダ番号、トラック番号とも０から開始される。図４における仮想ボリューム（１１００）データ領域の説明で、「シリンダｉのトラックｊ」とは、「シリンダ番号がｉで、かつ、ヘッダ番号がｊであるトラック」に対応するデータ領域であることを意味する。

【００３４】図４で、仮想ボリューム管理テーブル（１

４００）内のデータ先頭ポインタ（１４５０）の値は４０９６に設定されている。すなわち、本実施の形態では、仮想ボリューム（１１００）用のデータ領域を確保した先頭番地は、論理区画１（１０００）の実記憶装置上で４０９６番地であることを表す。

【００３５】図４で、アクセスポインタ（１４６０）の値は５２８３８４に設定されている。すなわち、次に仮想ボリューム（１１００）に対するアクセス要求が発生したときに、論理区画１（１０００）の実記憶装置上で５２８３８４番地に位置するデータを処理することを表す。ここで、５２８３８４番地とは、「データ先頭ポインタ（２４５０）＋仮想ヘッダ数（２４２０）×仮想トラック長（２４３０）」と一致し、仮想ボリューム（１１００）内のシリンダ１、トラック０に相当する。

【００３６】図５は、論理区画２（２０００）内のボリューム識別テーブル（２６００）の構成を示している。ボリューム識別テーブル（２６００）は、論理区画２（２０００）内で実行されるＡＰ（２１００）がアクセス可能なディスク装置の一覧を、ボリュームＩＤ（２６１０）とボリューム種別（２６２０）の組合せにより示している。この情報は、システム構成情報として、システム立上げ時のパラメタ等によって、利用者があらかじめ登録した情報を格納する。

【００３７】ボリュームＩＤ（２６１０）は、アクセス対象のディスクを識別するために利用する情報である。本実施の形態では、登録順に１、２．．．なる昇順の整数により、各ディスクを識別する。システムに登録された順序は、物理ディスク（４０００）、仮想ボリューム（１１００）であるとする。すなわち、ボリュームＩＤ（２６１０）＝１は物理ディスク（４０００）に、また、ボリュームＩＤ（２６１０）＝２は仮想ボリューム（１１００）に、それぞれ対応している。

【００３８】ボリューム種別（２６２０）は、ボリュームＩＤ（２６１０）で示されたディスク装置の種類が、物理ディスクか、あるいは仮想ボリュームであるかを識別するための情報を格納するフィールドである。本実施の形態では、ボリューム種別（２６２０）＝Ｐは物理ディスク、ボリューム種別（２６２０）＝Ｖは仮想ボリュームであることを表す。

【００３９】図６は、論理区画２（２０００）内のＡＰ（２１００）から要求された入出力動作を実行するために必要な情報を格納したチャネルコマンド語（以下ＣＣＷ）の構成を示している。５０００は、物理ディスク（４０００）に対して処理を要求するときに生成するチャネルコマンド語の構成を示している。ＣＣＷ（５０００）は、データ管理プログラム（２２００）が生成する。また、５１００は、仮想ボリューム（１１００）に対するアクセス要求時、論理区画１（１０００）との間の通信に利用される通信ＣＣＷの構成を示している。通信ＣＣＷ（５１００）は、データ管理プログラム（２２００）によって起動された仮想ボリュームアクセス要求プログラム（２３００）が生成する。

【００４０】ＣＣＷ（５０００）は、コマンド（５０１０）とデータアドレス（５０２０）を有する。コマンド（５０１０）は、入出力装置に対してどのような処理を要求するかを示す領域で、本実施の形態では、位置付け、書込み、読込みの三種類を指定できる。このうち、位置付けコマンドとは、ディスク装置内での位置情報（位置付け情報）を指定し、次に処理すべきデータ位置を指示するコマンドである。

【００４１】データアドレス（５０２０）は、処理対象データが存在する実記憶装置上でのアドレスを示す領域である。位置付け要求の場合は、前記した位置付け情報が存在するアドレスが、このデータアドレス（５０２０）で指定される。

【００４２】対象が仮想ボリューム（１１００）の場合は、ＣＣＷ（５０００）をもとにして、仮想ボリュームアクセス要求プログラム（２３００）が通信ＣＣＷ（５１００）を生成する。更に、仮想ボリューム（１１００）アクセス完了後、仮想ボリュームアクセス実行プログラム（１２００）からの応答時にも、この通信ＣＣＷ（５１００）が生成される。通信ＣＣＷ（５１００）は、コマンド（５１１０）、データアドレス（５１２０）、データ（５１３０）によって構成される。このうち、コマンド（５１１０）、データアドレス（５１２０）は、それぞれＣＣＷ（５０００）のコマンド（５０１０）、データアドレス（５０２０）と同じ意味を持つ。データ（５１３０）は、コマンド（５１１０）によって示された要求種別に従い、データ本体も転送する必要がある場合に、そのデータ本体を格納するための領域である。

【００４３】以下、仮想ボリューム（１１００）をアクセスするときのＣＣＷ（５０００）と通信ＣＣＷ（５１００）の生成状況を、位置付け要求、書込み要求、読込み要求の順に説明する。

【００４４】図７は、ＡＰ（２１００）から仮想ボリューム（１１００）に対して位置付け要求を行う際の論理区画２（２０００）側の状態を示す。ＣＣＷ（５０００）のデータアドレス（５０２０）に、位置付け情報（２１１０）のアドレスが格納される。この位置付け情報（２１１０）は、論理区画２（２０００）の実記憶装置内に存在する。そのため、論理区画１（１０００）から直接、内容を読み取ることはできない。そこで、論理区画１（１０００）に対する通信ＣＣＷ（５１００）のデータ（５１３０）に、データアドレス（５０２０）が示す位置付け情報（６０００）を格納し、通信ＣＣＷ（５１００）を論理区画１（１０００）に送信する。

【００４５】図８は、位置付け要求通信を受け取った論理区画１（１０００）側の状態を示す。論理区画１（１０００）側では、通信ＣＣＷ（５１００）内データ（５

１３０）に格納された位置付け情報をもとにして、仮想ボリューム管理テーブル（１４００）内のアクセスポインタ（１４６０）を更新する。元の位置付け情報は、物理ディスク装置内での配置場所を記述した形式となっているため、この情報を論理区画１（１０００）内で仮想ボリューム（１１００）を配置した実記憶装置内でのアドレスに変換する必要がある。この変換処理の詳細については、のちほど仮想ボリュームアクセス実行プログラム（１２００）の処理フローチャートを説明する際に、詳しく説明する。

【００４６】位置付け処理が完了したのち、論理区画２（２０００）に対する応答として、コマンド（５１１０）、データアドレス（５１２０）、データ（５１３０）の内容が要求時と同じ通信ＣＣＷ（５１００）を生成し、論理区画２（２０００）に通信する。

【００４７】引き続き、図９、図１０を用いて、仮想ボリューム（１１００）への書込み処理を行う際の状態を説明する。図８は、データの書込みを要求する論理区画２（２０００）側の状態を示す。ＣＣＷ（５０００）内のデータアドレス（５０２０）は、ＡＰ（２１００）から仮想ボリュームへ書き込むべきデータが格納された入出力バッファ（２１２０）のアドレスを示している。この入出力バッファ（２１２０）の内容をデータ（５１３０）に格納した通信ＣＣＷ（５１００）を生成し、論理区画１（１０００）に送信する。

【００４８】図１０は、書込み要求を受け取った論理区画１（１０００）側の状態を示している。通信ＣＣＷ（５１００）内データ（５１３０）に格納された内容を、仮想ボリューム（１１００）のうち、仮想ボリューム管理テーブル（１４００）内のアクセスポインタ（１４６０）によって示されるデータ領域（１１１０）に複写する。論理区画２（２０００）に対する応答の通信ＣＣＷ（５１００）では、データ（５１３０）に情報を格納する必要はない。

【００４９】引き続き、図１１から図１３を用いて、仮想ボリューム（１１００）への読込み要求を行う際の状態を説明する。

【００５０】図１１は、読込みを要求する論理区画２（２０００）側の状態を示す。論理区画１（１０００）に対して送信する通信ＣＣＷ（５１００）のうち、コマンド（５０００）、データアドレス（５１００）の内容はＣＣＷ（５０００）と同様である。データ（５１３０）には、アクセス要求時には情報を格納する必要はない。

【００５１】図１２は、読込み要求通信を受け取った論理区画１（１０００）側の状態を示している。アクセスポインタ（１４６０）が示す仮想ボリューム（１１００）内のデータ領域（１１１０）から、通信ＣＣＷ（５１００）内データ（５１３０）へデータを複写して、論理区画２（２０００）に通信ＣＣＷ（５１００）を応答する。

【００５２】図１３は、読込み応答通信を受け取った論理区画２（２０００）側の状態を示している。通信ＣＣＷ（５１００）内データ（５１３０）に格納された内容、すなわち仮想ボリューム（１１００）から読み込んだデータを、データアドレス（５１２０）が示すＡＰ（２１００）内の入出力バッファ（２１２０）に複写する。

【００５３】図１４は、仮想ボリューム（１１００）へのアクセス要求を受け取ったときに実行される論理区画１（１０００）内の仮想ボリュームアクセス実行プログラム（１２００）の処理フローチャートである。

【００５４】まず、応答通信の準備として、要求された通信ＣＣＷ（５１００）内のコマンド（５１１０）、データアドレス（５１２０）を、応答通信用の通信ＣＣＷ（５１００）のコマンド（５１１０）、データアドレス（５１２０）に格納（ステップ１２１０）する。

【００５５】次に、要求されたコマンド種別を判定（ステップ１２２０）する。ここで、コマンド種別には、位置付け、書込み、読込みの三種類が存在する。コマンド種別が位置付け要求だった場合は、通信ＣＣＷ（５１００）内データ（５２００）に格納されている位置付け情報を、仮想ボリューム（１１００）を格納した論理区画１（１０００）内の実記憶装置のアドレスに変換（ステップ１２３０）する。位置付け情報として指示される指定は、物理ディスク（４０００）をアクセスするための情報と共通であり、「ＣＣＨＨＲＲ」という形式で示される。ここで、ＣＣはシリンダ番号、ＨＨはヘッダ番号、Ｒはレコード番号である。このような指定に対し、仮想ボリューム（１１００）内のデータ配置先として、「データ先頭ポインタ（２４５０）＋（ＣＣ×仮想ヘッダ数（２４２０）＋ＨＨ）×仮想トラック長（２４３０）＋Ｒ×レコード長（２４４０）」なる計算を実行することによって、仮想ボリューム（１１００）内データ領域のアドレスに変換する。更に、変換したデータを、アクセスポインタ（１４６０）に格納（ステップ１２４０）する。

【００５６】コマンド種別が書込み要求だった場合は、通信ＣＣＷ（５１００）内データ（５１３０）の内容を、アクセスポインタ（１４６０）が示す仮想ボリューム（１１００）内のデータ領域に複写（ステップ１２５０）する。また、コマンド種別が読込み要求だった場合は、アクセスポインタ（１４６０）が示す仮想ボリューム（１１００）内のデータ領域から、応答通信用の通信ＣＣＷ（５１００）内データ（５１３０）へ内容を複写（ステップ１２６０）する。

【００５７】最後に、以上のステップで作成した応答通信用の通信ＣＣＷ（５１００）を、要求元の論理区画２（２０００）に対して送信することを、通信プログラム（１５００）に依頼（ステップ１２７０）する。

【００５８】図１５は、論理区画２（２０００）において、ＡＰ（２１００）から入出力要求を受け取ったときに実行されるデータ管理プログラム（２２００）の処理フローチャートである。

【００５９】はじめに、入出力装置に対して要求を実行するためのＣＣＷを作成（ステップ２２１０）する。次に、入出力要求を実行する装置のボリュームＩＤを取得（ステップ２２２０）する。引き続き、アクセス対象入出力装置が、仮想ボリュームか、あるいは物理ディスクかを判定（ステップ２２３０）する。この判定処理は、ボリューム識別テーブル（２６００）を参照し、ステップ２２２０で取得したボリュームＩＤ（２６１０）に対応するボリューム種別（２６２０）をチェックすることによって実現できる。

【００６０】アクセス対象が、物理ディスクであった場合、該当する入出力装置に対して入出力起動命令を実行（ステップ２２４０）する。この物理ディスク装置へアクセスする場合の処理は、論理区画１（１０００）内に仮想ボリューム（１１００）を生成しない計算機システムと同様である。

【００６１】ステップ２２３０の判定において、アクセス対象が仮想ボリューム（１１００）だった場合、仮想ボリューム（１１００）が存在する論理区画１（１０００）への通信を実行するため、仮想ボリュームアクセス要求プログラム（２３００）を実行（ステップ２２５０）する。

【００６２】図１６は、仮想ボリュームアクセス要求プログラム（２３００）の処理フローチャートである。

【００６３】通信ＣＣＷ（５１００）を構成するコマンド（５１１０）、データアドレス（５１２０）に、それぞれデータ管理プログラム（２２００）が作成したＣＣＷ（５０００）内のコマンド（５０１０）、データアドレス（５０２０）と同様の情報を格納（ステップ２３１０）する。次に、コマンド種別を判定（ステップ２３２０）する。判定の結果、コマンド種別が位置付け、または書込みだった場合、通信ＣＣＷ（５１００）のデータ（５１３０）に、データアドレス（５１２０）が示す領域の内容を格納（ステップ２３３０）する。この情報は、位置付け要求の場合は位置付け情報（６０００）であり、書込み要求の場合は仮想ボリューム（１１００）に書き込むべき入出力バッファ（７０００）内のデータである。

【００６４】以上の処理ステップによって作成した通信ＣＣＷ（５１００）を、仮想ボリューム（１１００）が存在する論理区画１（１０００）に対して送信することを、通信プログラム（２５００）に依頼（ステップ２３４０）する。

【００６５】図１７は、仮想ボリューム（１１００）へのアクセスが完了したとき、論理区画１（１０００）から応答された通信ＣＣＷ（５１００）を受け取った通信プログラム（２５００）が起動する仮想ボリュームアクセス完了プログラム（２４００）の処理フローチャートである。

【００６６】まず通信ＣＣＷ（５１００）の内容から、コマンド種別を判定（ステップ２４１０）する。同判定において、コマンド種別が読込みだった場合、通信ＣＣＷ（５１００）内データ（５１３０）の内容を、データアドレス（５１２０）が示す実記憶装置内の領域、すなわちＡＰ（２１００）が指定した入出力バッファ（７０００）へ複写（２４２０）する。ステップ２４１０の判定において、コマンド種別が位置付け、または書込みであった場合は、通信ＣＣＷ（５１００）からのデータ複写は必要ない。

【００６７】続けて、入出力要求元のＡＰ（２１００）を再起動するために、データ管理プログラム（２２００）を実行（ステップ２４３０）する。

【００６８】次に、図１８を用いて、本発明を用いた第２の実施の形態を説明する。

【００６９】第２の実施の形態では、論理区画１（１０００）、論理区画２（２０００）に加え、論理区画３（６０００）と、論理区画３（６０００）に割当てられた物理ディスク（７０００）が存在する。論理区画３（６０００）の構成内容は論理区画２（２０００）と同様であり、説明は省略する。このような構成を取ることによって、論理区画２（２０００）で実行されるＡＰ（２１００）と、論理区画３（６０００）で実行されるＡＰ（６１００）が、仮想ボリューム（１１００）内のデータを共用することができる。

【００７０】次に、図１９を用いて、本発明を用いた第３の実施の形態を説明する。

【００７１】第３の実施の形態における計算機システムは、１つの計算機を論理区画に分割して実行するシステムではなく、複数の計算機ノードから成るクラスタ型計算機システムである。ここで、計算機システムは、計算機ノード１（８０００）と、計算機ノード２（９０００）とから構成される。また、１００００は、計算機ノード２（９０００）に接続された物理ディスク装置であり、１１０００は、計算機ノード１（８０００）と計算機ノード２（９０００）の間でデータ通信を実行するための通信経路である。

【００７２】計算機ノード１（８０００）内の構成内容は、第１の実施の形態における論理区画１（１０００）と同様である。同じように、計算機ノード２（９０００）内の構成内容は、第１の実施の形態における論理区画２（２０００）と同様である。仮想ボリューム（８１００）をアクセスするために、計算機ノード間で通信するデータの形式も、第１の実施の形態で説明した通信ＣＣＷ（５１００）と同じ構成とする。これにより、クラスタ型計算機システムにおいても、計算機ノード１（８０００）の実記憶装置上に生成した仮想ボリューム（８

１００）を、計算機ノード２（９０００）のＡＰ（９１００）から利用することが可能となる。更に計算機ノード数を増加した構成では、仮想ボリューム（８１００）内のデータを、複数の計算機ノード間で共用することができる。

【００７３】

【発明の効果】本発明によれば、共用するデータを外部記憶装置に配置していたシステムで動作していたプログラムを変更せずに、共用データへのアクセスの高速化を図ることができるので、システムの応答性能やスループットを向上できる。また、本発明によれば、共用しないデータに関しても、メモリを利用することで、アクセスの高速化を図り、システムの応答性能やスループットを向上できる。また、本発明によれば、こうしたデータをへのアクセスに関して、別の論理区画域に配置させることで、データの機密性を向上させることができる。

【図面の簡単な説明】

【図１】　第１の実施の形態を示す計算機システムの構成図

【図２】　第１の実施の形態における制御の流れ

【図３】　第１の実施の形態における仮想ボリューム管理テーブルの構成図

【図４】　第１の実施の形態における仮想ボリューム初期設定の状態

【図５】　第１の実施の形態におけるボリューム識別テーブルの構成図

【図６】　第１の実施の形態におけるＣＣＷと通信ＣＣＷの構成図

【図７】　第１の実施の形態における位置付け要求時の論理区画２の状態

【図８】　第１の実施の形態における位置付け要求時の論理区画１の状態

【図９】　第１の実施の形態における書込み要求時の論理区画２の状態

【図１０】　第１の実施の形態における書込み要求時の論理区画１の状態

【図１１】　第１の実施の形態における読込み要求時の論理区画２の状態

【図１２】　第１の実施の形態における読込み要求時の論理区画１の状態

【図１３】　第１の実施の形態における読込み要求時の論理区画２の状態

【図１４】第１の実施の形態における仮想ボリュームアクセス実行プログラムの処理フローチャート

【図１５】第１の実施の形態におけるデータ管理プログラムの処理フローチャート

【図１６】第１の実施の形態における仮想ボリュームアクセス要求プログラムの処理フローチャート

【図１７】第１の実施の形態における仮想ボリュームアクセス完了プログラムの処理フローチャート

【図１８】第２の実施の形態を示す計算機システムの構成図

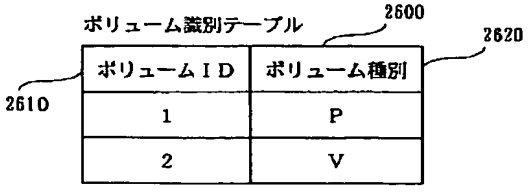
【図１９】第３の実施の形態を示す計算機システムの構成図

【符号の説明】

１０００……論理区画１、１１００……仮想ボリューム、１２００……仮想ボリュームアクセス実行プログラム、１３００……仮想ボリューム生成プログラム、１４００……仮想ボリューム制御テーブル、１５００……通信プログラム、２０００……論理区画２、２１００……アプリケーションプログラム、２１１０……位置付け情報、２１２０……入出力バッファ、２２００……データ管理プログラム、２３００……仮想ボリュームアクセス要求プログラム、２４００……仮想ボリュームアクセス完了プログラム、２５００……通信プログラム、３０００……ハイパバイザ、３１００……計算機資源配分プログラム、３２００……論理区画間通信プログラム、４０００……物理ディスク

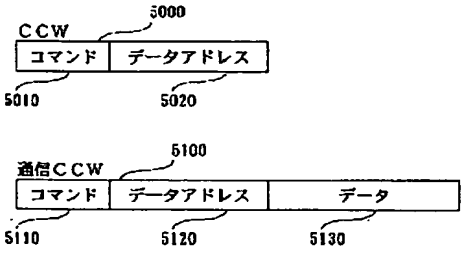
【図５】

図５



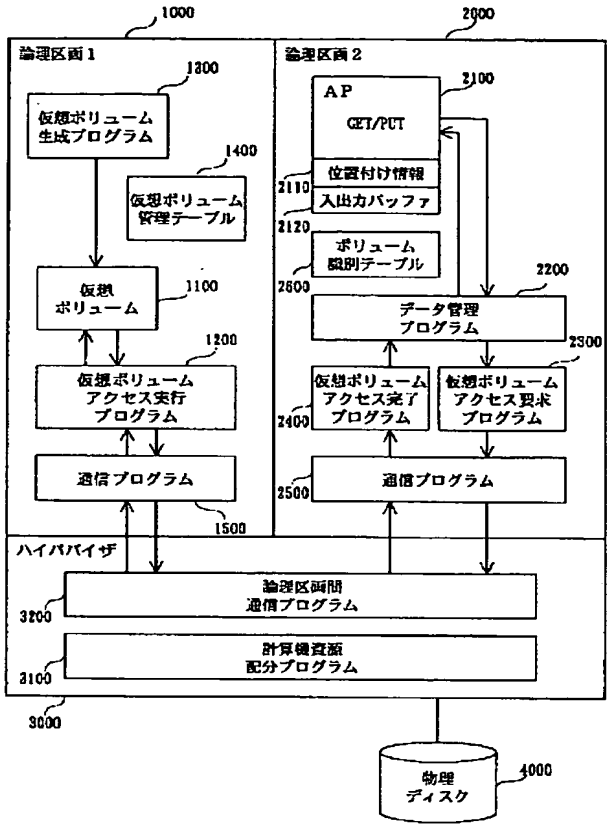
【図６】

図６



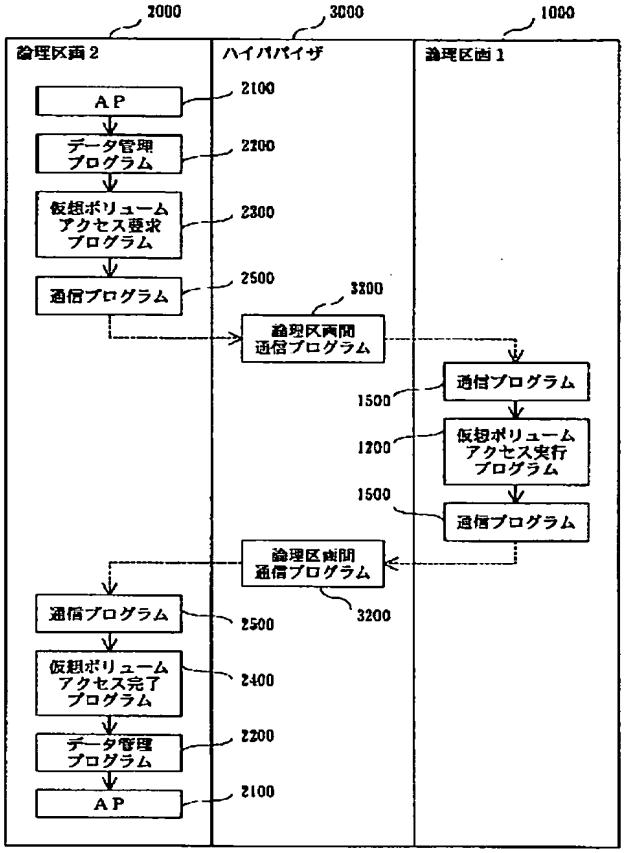
【図１】

図１



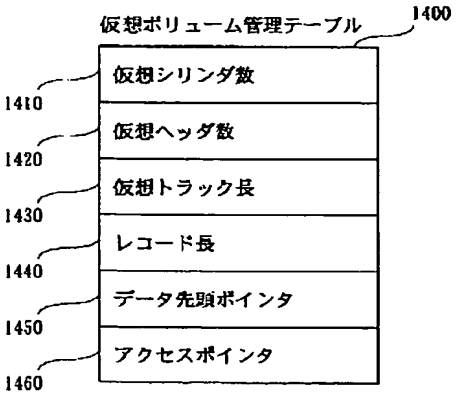
【図２】

図２



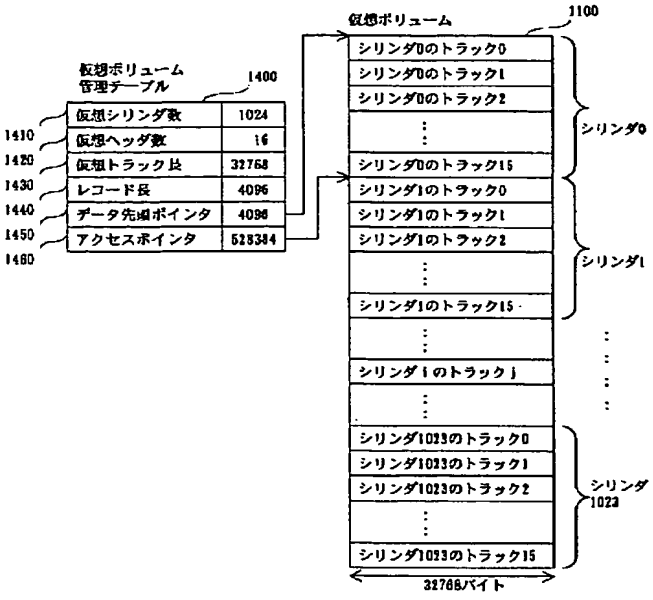
【図３】

図３



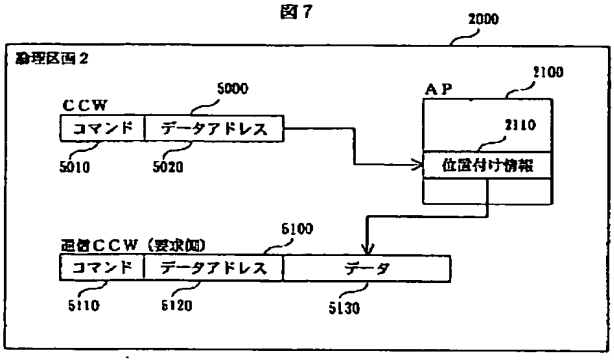
【図４】

図４



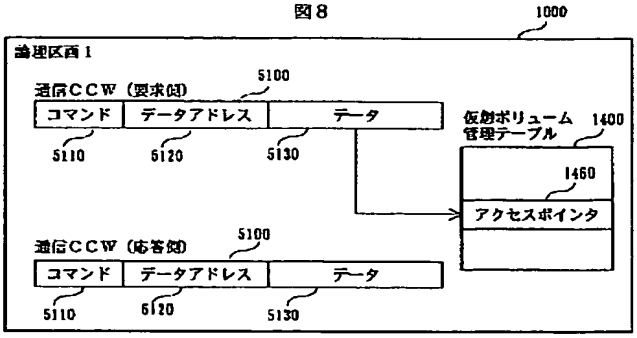
【図７】

図７



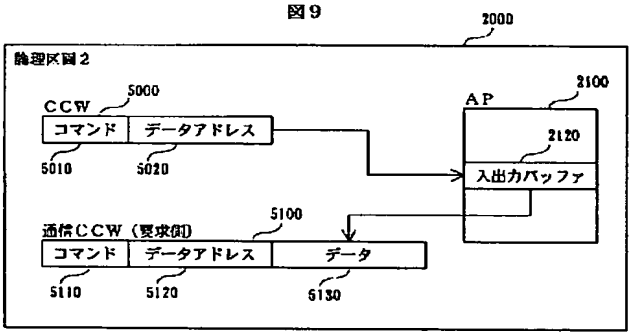
【図８】

図８



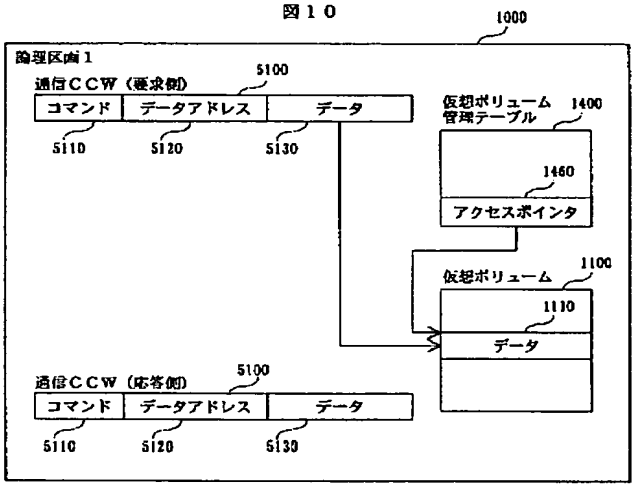
【図９】

図９



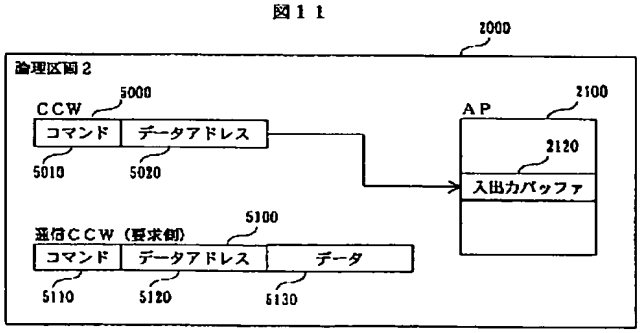
【図１０】

図１０



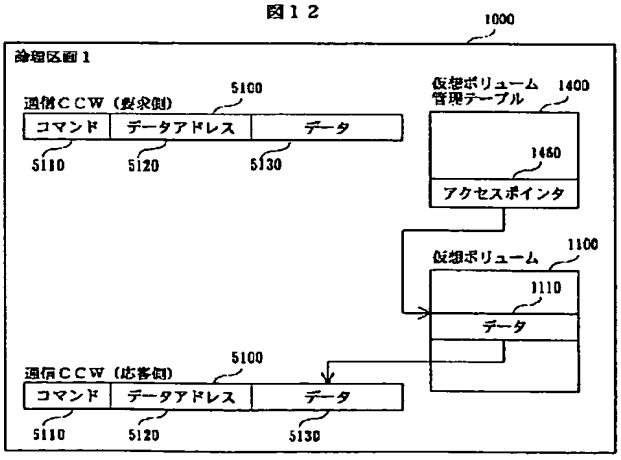
【図１１】

図１１



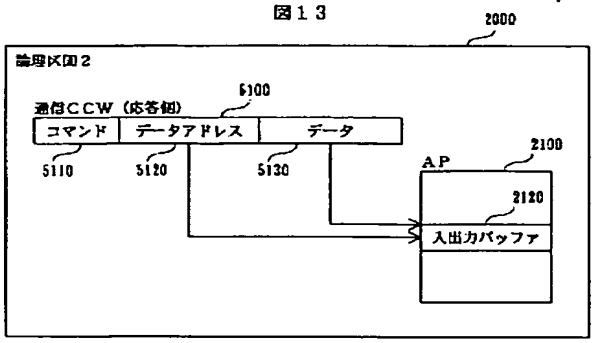
【図１２】

図１２



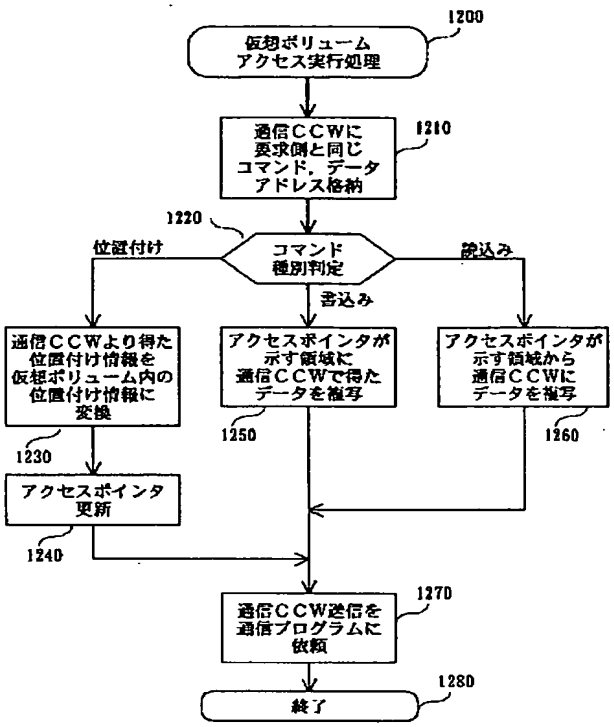
【図１３】

図１３



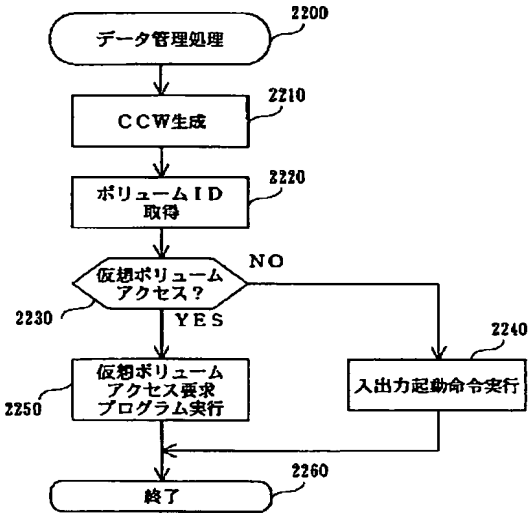
【図１４】

図１４



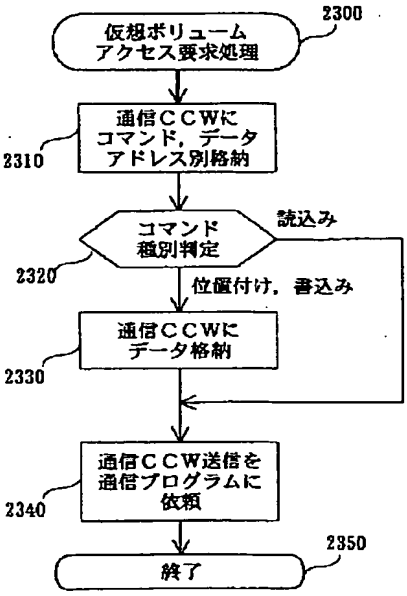
【図１５】

図１５



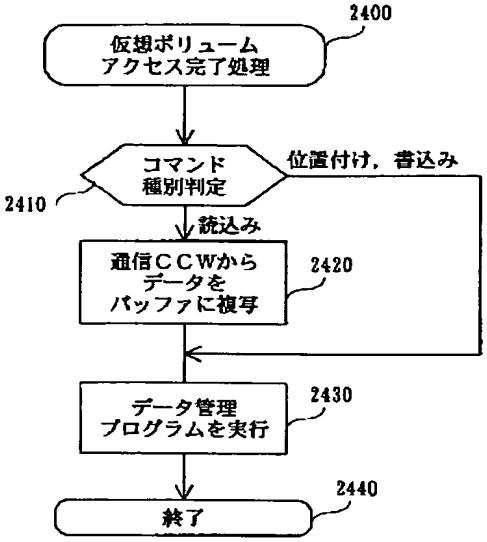
【図１６】

図１６



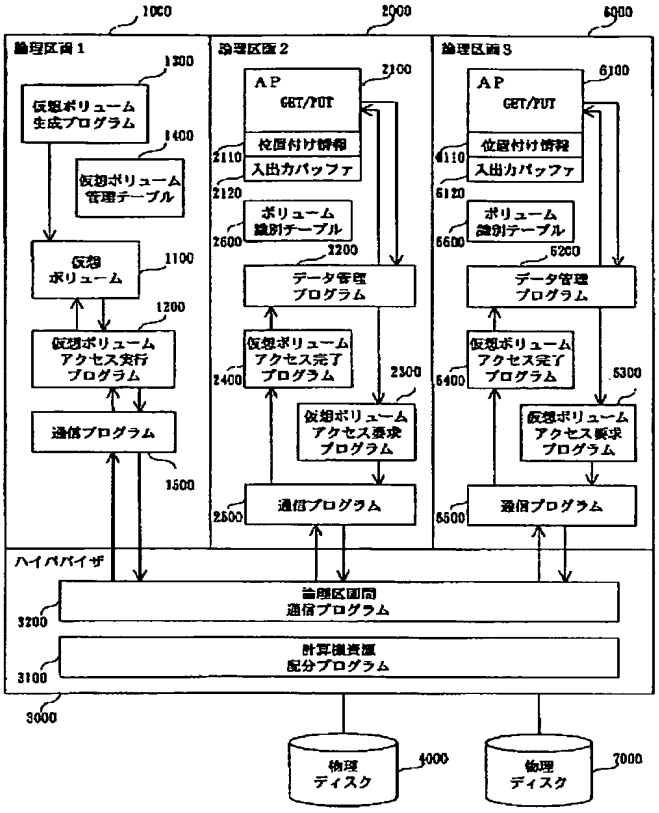
【図１７】

図１７



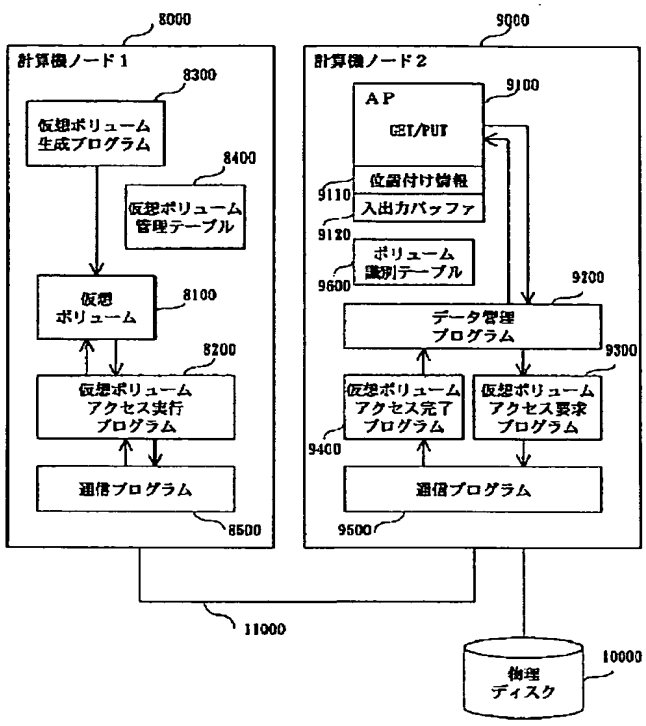
【図１８】

図１８



【図１９】

図１９



フロントページの続き

(72)発明者　大辻　彰
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウエア事業部内
(72)発明者　池ケ谷　直子
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内
(72)発明者　平岩　友理
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

Ｆターム(参考)　5B065 BA01 CA02 CA13 CC02 CC08 CH04 PA12
5B098 AA03 GA01 GD03 GD15 HH04 HH07